Discovering Relations among Named Entities by Detecting Community Structure

Tingting He[1 2] ,Junzhe Zhao[1] ,Jing Li[1]
[1] Department of Computer Science, Huazhong Normal University 430079 Wuhan, China
[2] Software College of Tsinghua University 102201 Beijing, China
tthe@mail.ccnu.edu.cn zhaojunzhe@mails.ccnu.edu.cn lee_king8@hotmail.com

Abstract. This paper proposes a networked data mining method for relations discovery from large corpus. The key idea is representing the named entities pairs and their contexts as the network structure and detecting the communities from the network. Then each community relates to a relation the named entities pairs in the same community have the same relation. Finally, we labeled the relations. Our experiment using the corpus of People's Daily reveals not only that the relations among named entities could be detected with high precision, but also that appropriate labels could be automatically provided for the relations.

Keywords: named entities pair, community, betweenness

1 Introduction

Relation extraction among named entities (NR) is one of major tasks in information extraction (IE).The goal of relation extraction is to find out the relations among named entities(NE) in documents. In recent years, such technology has been widely used in many fields, such as: information retrieval, question-answering systems, biology technology, construction of ontology, etc.

Many methods have been proposed for relation extraction, including supervised learning methods (Zelenko et al., 2002) [8], weakly supervised learning methods (Brin 1998; Agichtein 2000; Sudo 2003) [9][10][11] and unsupervised learning methods (Hasegawa 2004; Chen Jinxiu 2005) [1][2]. In this paper, we propose an unsupervised method for relation extraction.

Currently, unsupervised learning methods for relation extraction have some difficulties. For example, in Hasegawa et al.'s method, they eliminated less frequent NE pairs, collected the contexts of NE pairs, clustered the contexts using complete linkage method, and finally selected the most frequent word in a cluster to label the relation in this cluster. However, the less frequent NE pairs might have relations, and it is difficult for complete linkage method to select the threshold.

In Chen Jinxiu et al.'s method, they firstly got the NE pairs labeled with relation in ACE corpus and collected their contexts, clustered the contexts using stability-based method, and finally used DCM method to label the clusters. However, we can't select NE pairs with relation in the unannotated corpus. And the relations of NE pairs might be in hierarchical structure; but the stability-based method can not discover the hierarchical structure.

Based on the issues mentioned above, this paper proposes a method to discover relations among NEs based on networked data mining. The advantages of this method is that there is no need to eliminate less frequent NE pairs and select the NE pairs which have relations. In addition, our method can automatically present the relations of NE pairs in the hierarchical structure.

The rest of this paper is organized as follows. Section 2 talks about the construction for networked structure. Section 3 explores the procedure of relation discovering and labeling. Section 4 describes experiments and evaluation of experiment results. Section 5 gives a discussion of the problems in existing methods. Finally, Section 6 gives a summary and talks about future prospect.

2 Network constructing

Many systems take the form of networks. Sets of nodes or vertices joined together in pairs by links or edges. Researches show that the distribution of these network vertices forms many communities, which means vertices are classified into different groups; there are many edges connecting the vertices inside a same group and few connections between groups. However, vertices in a same group could be divided into smaller, tighter structured groups. We call this group community [3] [13]. So we construct a network by NE pairs and their contexts for relations discovery.

2.1 NE pairs and their contexts

We define NE pair as bellow: In the corpus, if there is a sentence s_i which contain 2 named entities e_x,e_y; and the number of words between e_x,e_y is no more than N, then e_x,e_y make a NE pair.

Firstly, we have to define a window of context (WIN*pre*-WIN*mid*-WIN*post*). For a random selected NE pair (e_x,e_y), we get all the sentences which includes this entities pair; then in each sentence, retrieve WIN*pre* words in front of e_x, WIN*post* words after e_y, and all words between e_x and e_y; use them as the context of the sentence; Finally, add all the contexts of the sentence to a group, the group will be the context of NE pair (e_x,e_y).

2.2 Construction for network

We present NE pair as weighted network structure: NE pairs are vertices; if the corresponding contexts of 2 NE pairs have common words, we connect the corresponding network vertices with an edge, the weight of the edge is defined as the similarity of the corresponding contexts of connected NE pairs [14], as in equation (1):

$$Weight_{ij} = \frac{C(context_i, context_j)}{context_i + context_j} = \frac{\sum_{k=1}^{m} \min(nw_{ki}, nw_{kj})}{\sum_{k=1}^{m} (nw_{ki} + nw_{kj})} \quad (1)$$

Where $Weight_{ij}$ is the similarity between contexts of NE pair *i* and *j*, *m* is the number of words in $context_i$ and $context_j$. nw_{ki} is the times that word *k* appears in the $context_i$.

3 Relations discovery and labeling

3.1 Betweenness clustering

Researches show that the edges between communities have great betweenness. Thus, we just cut the edges with great betweenness to get expected communities [7]. In general, the betweenness of edge is defined as the accumulated times for shortest path to pass such edge between network vertices. In terms of the weight network in this paper, we modify the betweenness of edge *e* as below (2):

$$Betweenness(e) = \frac{\sum between(v_1, v_2)}{e.W} \quad (2)$$

Where $\sum_{between\ (v_1, v_2)}$ is the accumulated times for shortest path to pass edge e between network vertices; $e.W$ is the weight of edge e.

In order to find expected communities in network N, we should know when to stop dividing network N. So when N has been divided into g communities, we need to evaluate the modularity in divided result [4], as below (3):

$$Q = \sum_i (e_{ii} - (\sum_j e_{ij})^2) = Tr(e) - \left\| e^2 \right\| \quad (3)$$

Let us define a $g \times g$ matrix e whose element e_{ij} is the fraction of all edges in the network that link vertices in community i to vertices in community j. We modify e_{ij} because of weighted network, as in equation (4); $Tr(e)=\sum_i e_{ii}$ gives the fraction of edges in the network that connect vertices in the same community; $\| e \|$ indicates the sum of the elements of the matrix e.

$$e_{ij} = \frac{\sum e.W}{\sum e'.W} \quad (4)$$

Where e is an edge connecting community i and j; e' is an edge of network N. We have to notice that: the edges connecting community i and j are defined as all the edges in community $i(j)$, if community i and j are the same community.

When we get the best part evaluation result, we recognize that each community relates to a relation. Table 1 is the algorithm implementation of automatically discovering the communities from network. We can discover hierarchical community structure by repeatedly executing the algorithm [12][15].

Table 1. Automatically discovering communities

	Input: N(initial network);
	Output: S_N(the community set belong N);
1	initialize S_N={N};
2	initialize S_N modularity Q=0;
3	choose a network N_i which is never divided in S_N, calculate the betweenness of each edge in N_i, ceaselessly cut the edge with greatest betweenness until divide N_i to community $N_{i1}$and N_{i2};
4	calculate S_N modularity Q', if $Q'>Q$ or $Q=Q'$, move out N_i from S_N, and add N_{i1} and N_{i2} to S_N;
5	if exist the network which is never divided in S_N, then to step 2;else to step 6;
6	end and return S_N;

3.2 Relation description

In the hierarchical structure relations of NE pairs, the relations in the bottom level are detailed; it could be described by a word in context. This paper uses the method of DCM (Chen Jinxiu, 2005) [2] to set the weight of every word, considering the importance of the word in a certain community and all communities. Then we label the NE pairs' relations with the word with the highest weight. Furthermore, we find that many communities just contain 1 NE pair in clustering result. And we find out a majority of these NE pairs without relation. So we collect these to be a set of NE pairs without relation.

4 Experiment and evaluation

4.1 Manual annotation

We quote 2 types of NE pair in annotated corpus of People's Daily (7 million characters) (Location-Person, Organization-Organization) as the research objects to verify the method in this paper. We get 1112 NE pairs (Loc-Per, Org-Org) from corpus and annotate relations to these NE pairs manually. Table 2 is the 2 leveled hierarchical relations.

Table 2. Manual annotation

Type of NE pair # Number of relations	Relation in 1st layer # Number of relations	Relation in 2nd layer # Number of instances
Loc- Per 18	国家-领导人 12	总统 总理 首相 … 47 29 14
	地方-行政官员 22	市长 省长 书记 … 36 21 44
	地方-记者 国家-运动员	None relation in second layer
	… 291 NE pairs without relation in all 845 NE pairs	
Org-Org 7	机构-合作机构 11	联合 合作 合资 … 45 20 14
	球队-球队 6	战胜 比赛 平局 … 38 21 19
	… 43 NE pairs without relation in all 267 NE pairs	

4.2 Experiment parameters and experiment result

During the process of the experiment, we select 2-6-2 sized window to get the contexts of NE pairs. In order to make evaluation simpler, we select 2 levels to be the highest deepness of hierarchical structure in the experiment, we make it the same deepness as manual work. Finally, we describe relations with words in context for 2nd level. And we manually summarize appropriate relations description for 1st level. Table 3 gives the number of communities in 2 levels in hierarchical structure from experiment.

Table 3. Compare of manual result and experimental result

Type of NE pair	Number of relations in 1st layer		Number of relations in 2nd layer	
	Manual work	Experiment	Manual work	Experiment
Loc-Per	18	19	49	43
Org-Org	7	9	27	31

4.3 Analysis and improvement

To analyze the result, we find out that there are some relations' descriptions in 2nd level with almost the same meaning. E.g.总统-总书记, 外交部长-外相, etc. So we apply the method and software

package from<Word Similarity Computing Based on How-net>（Qun Liu, 2002）[1] [16] to calculate the similarity between relations descriptive words in 2nd level. According to the result, groups like {总统, 总书记},{外交部长,外相},{战胜,击败,赢} have 1.0 point of similarity between internal words, which means semantically the same. So we think the relations these words described are the same. Thus we combine these relations, and use the aggregation of these words to describe 2nd level relations after combination. Finally we find out improved relation description is more reasonable, and more logical in real life application. Improved relation description is in Table 4.

Table 4. The result of the improvement

Type of NE pair	Relations in 1st layer	Relations in 2nd layer
Loc-Pre	国家-领导人	{总统, 总书记}, {外交部长, 外相}, 总理, 首相 . . .
	地方-行政官员	市长, 省长, 书记, 镇长 . . .
	地方-记者 国家-运动员	None relation in second layer
	…	…
Org-Org	机构-合作机构	联合, 合作, 合资, 兼并 . . .
	球队-球队	{战胜, 击败, 赢}, 比赛, 平局 . . .
	…	…

4.4 Evaluation of the result

In order to quantify the consistency level between experimental result and manual result, we adopted a permutation procedure to assign different relation type tags to only min(|EC|,|TC|) clusters, where |EC| is the estimated number of clusters, and |TC| is the number of ground truth classes (relation types). This procedure aims to find a one-to-one mapping function Ω from the TC to EC which is based on the assumption that for any two clusters, they do not share the same class labels. Under this assumption, there are at most |TC| clusters which are assigned relation type tags. If the number of the estimated clusters is less than the number of the ground truth clusters, empty clusters should be added so that |EC| = |TC| and the one-to-one mapping can be performed [2]; it is shown in equation (5). And the mapping procedure can be formulated as the equation (6). We also use our data and our evaluated method to evaluate Hasegawa's method and Chen Jinxiu's method. Final compare is in Table 5.

$$\hat{\Omega} = \arg\max_{\Omega} \sum_{j=1}^{|TC|} T_{\Omega(j),j} \tag{5}$$

In equation (5), represent a NE pair community from experiment, it relates to No. *j* NEs relation group of manual work.

$$Accuracy(P) = \frac{\sum_j T_{\hat{\Omega}(j),j}}{\sum_{i,j} T_{i,j}} \tag{6}$$

Equation (6) shows the accuracy ratio between NE pair in experiment.

Table 4. Evaluation result

Method	Precision in 1st layer		Precision in 2nd layer		Precision in NE pairs without relation	
	Loc-Pre	Org-Org	Loc-Pre	Org-Org	Loc-Pre	Org-Org

[1].Download of the tool to calculate the similarity between two words：http://www.keenage.com/html/e_index.html

Hasegawa	none	none	76.62%	62.40%	none	none
Chen	none	none	80.37%	68.39%	none	none
Our	81.33%	72.22%	91.15%	70.63%	77.25%	60.92%

There are 34.4% NE pairs in Loc-Pre and 16.1% NE pairs in Org-Org without relation. And we can recognize a majority of NE pairs without relation in our method. So our method has more advances in Loc-Pre.

5 Discussion

The automatically discovering method of NE pair relation in this paper has achieved good experiment result. We still hope to discuss some problems and related solutions from following aspects.

- Context window
- Size of corpus

Too big a window will bring much noise, which makes the edge among communities ambiguous and cannot get the good result. Too small a window will lose a lot useful information, which makes internal relation of community not tight enough, and will lose many important relations of NE pairs. This paper uses experience value when setup the size of context window. How to achieve best window size still need further research.

This experiment uses the corpus of People's Daily as research object. But the corpus is limited, so we can carry out the experiment based on Girvan and Newman（2003）'s method [4]. However, if the corpus is larger and the amount of retrieved NE pairs increase, we should apply A. Clauset（2004）'s method [6], in order to quickly find communities in larger scope network.

6 Conclusion and future work

This paper presents a method of automatically discovering relations among NEs based on networked data mining. This method contains the advantages in Hasegawa and Chen Jinxiu's methods. As compare with their two methods, our method has the following features: it can retrieve the hierarchical structure relations in NE pairs with higher accuracy; it needn't eliminate less frequent NE pairs; it can find out the NE pairs without relation; and it can combine the relation with the same semantic meaning. But this method has some space to improve in future work. Firstly we will try to verify such method based on bigger scope and specialized corpus. Secondly, we need to find out a better method to evaluate the hierarchical structure result and a better method to evaluate relation labeling.

Reference

1. Takaaki Hasegawa, Satoshi Sekine, Ralph Grishman: Discovering Relations among Named Entities from Large Corpora, Proceeding of Conference ACL2004. Barcelona, Spain. (2004)
2. Chen Jinxiu, Ji Donghong, Tan Chew Lim, Niu Zhengyu: Automatic Relation Extraction with Model Order Selection and Discriminative Label Identification. 2nd International Joint Conference on Natural Language Processing (IJCNLP05). Jeju Island. Republic of Korea. (2005)
3. M. Girvan, M. E. J. Newman: Community structure in social and biological networks. Proc. Natl. Acad. Sci.USA99. (2002) 7821-7826
4. M. E. J. Newman, M. Girvan: Finding and evaluating community structure in networks. Preprint condmat/0308217 (2003)
5. Defense Advanced Research Projects Agency: Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann Publishers, Inc (1995)
6. A.Clauset, M. E. J. Newman, C. Moore: Finding Community Structure in Very Large Networks." Phys. Rev. E 70, 066111 (2004)
7. M. E. J. Newman: Detecting community structure in networks. J. B 38. (2004) 321-330

8. Dmitry Zelenko, Chinatsu Aone, Anthony Richardella: Kernel methods for relation extraction. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002) (2002) 71-78
9. Sergey Brin: Extracting patterns and relations from World Wide Web. In Proc. Of WebDB Workshop at 6th International Conference on Extending Database Technology (WebDB'98). (1998) 172–183
10. Eugene Agichtein, Luis Gravano: Extracting relations from large plain-text collections. In Proc. of the 5th ACM International Conference on Digital Libraries (ACM DL'00) (2000)85-94
11. Kiyoshi Sudo, Satoshi Sekine, Ralph Grishman: An Improved Extraction Pattern Representation Model for Automatic IE Pattern Acquisition. Proceedings of ACL 2003.Sapporo. Japan. (2003)
12. Satoshi Sekine, Kiyoshi Sudo, Chikashi Nobata: Extended named entity hierarchy. In Proc. of the Third International Conference on Language Resources and Evaluation (LREC-2002) (2002)1818-1824
13. D. Gibson, J. Kleinberg, P. Raghavan: Inferring web communities from link topology. In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia. Association of Computing Machinery. New York (1998)
14. G. W. Flake, S. R. Lawrence, C. L. Giles, F. M. Coetzee: Self-organization and identification of Web communities. IEEE Computer 35 (2002) 66-71
15. P. Holme, M. Huss: Discovery and analysis of biochemical subnetwork hierarchies. Preprint qbio.MN/0309011 (2003)
16. Qun Liu, Jiansu Li:Word Similarity Computing Based on How-net(2002)